Anderson, Christopher

PREDICT 456 Section 55

Assignment #1

Introduction

With an emphasis throughout professional sports on "bigger, stronger, and faster" performers, today's professional athletes are physically larger than ever. This is evident in the NFL, where high-speed collisions become increasingly violent each year, and in the NBA, where the game moves at a faster pace than ever before. In Major League Baseball (MLB), while the paradigm has shifted from the muscle-bound players of the Steroid Era, teams search for pitchers who are tall, strong, throw extremely hard. The average height of an American male is approximately 5-foot-10, yet as of April 2015, 14 MLB teams did not have a single pitcher under 6 feet tall on their roster (Bryant). In the 2016 first year player draft, the Chicago Cubs selected five pitchers who are taller than 6'4" out of their first eight picks (Chicago Cubs 2016 Draft Results).

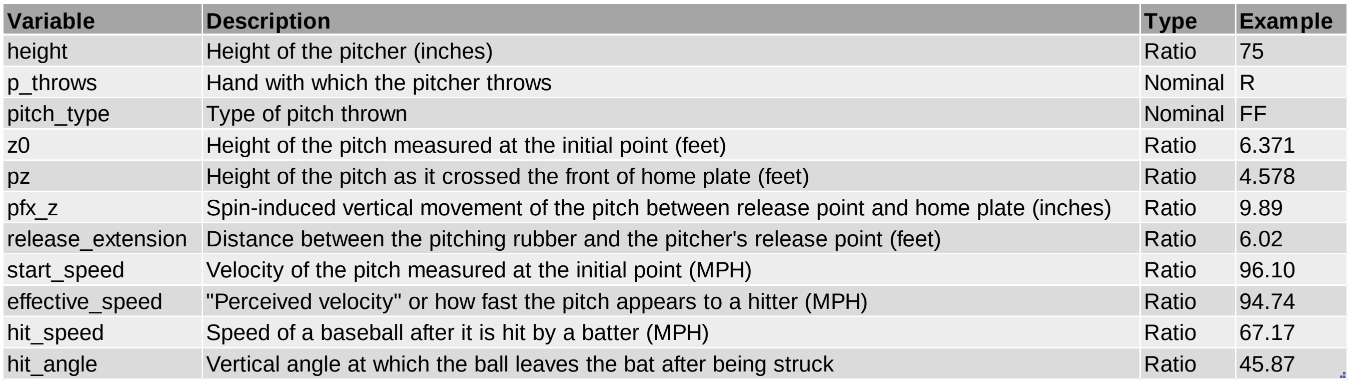
Logically, it makes sense that taller pitchers might throw with greater velocity because they are typically bigger and stronger. This was examined by baseball writer Zachary D. Rymer is 2013, with mixed results that did not show with certainty that taller pitchers throw harder. It also makes sense physiologically. They are able to extend farther towards home plate before releasing the pitch, meaning the batter has less time to react while the ball travels. In other words, taller pitchers may *appear* to throw the ball harder than they actually do, creating a "perceived velocity" greater than the actual velocity. Furthermore, taller pitchers can release the ball from a higher point than others, creating a more dramatic downward plane. Generally, pitches become more difficult to hit with greater vertical or horizontal movement. This trend towards taller pitchers was studied in detail in 2010 by Glenn Greenberg, and the results showed no significant difference between taller and shorter pitchers in effectiveness or durability (Greenberg).

Greenberg's study used traditional pitching statistics (e.g. Innings Pitched, Earned Run Average, etc.), and there are more advanced metrics available today that may shed more light on the topic. The purpose of this report is to assess the validity of the theory that there is a positive correlation between height and pitching effectiveness in the current era, using data from the PITCHf/x motion-tracking system. Created in the early 2000s and now used in every MLB stadium, PITCHf/x captures data on the release point, speed, and trajectory of every pitch that occurs in every game throughout the league (Marchi, 21). It also captures data on the speed and trajectory of the ball after it is put in play by the batter, which provide a new way to evaluate pitching and hitting effectiveness in the form of "exit velocity" and "launch angle" (Glossary). The hypotheses are: 1) there will be a positive correlation between height and perceived velocity; 2) there will be a positive correlation between height and the vertical movement, or downward plane, of the pitch; 3) there will be negative correlations between perceived velocity and both exit velocity and launch angle; and 4) there will be negative correlations between downward plane and both exit velocity and launch angle, supporting the anecdotal evidence in favor of MLB's preference for taller pitchers.

Methods

PITCHf/x data for every MLB game during the 2016 season through the end of June was extracted from the Statcast Search page on MLB's website. This resulted in a dataset of 348,728 observations across 60 variables, spanning April, May and June 2016. This investigation relates to the current era of Major League Baseball, so historical data was ignored. Including historical data may have also made the dataset unmanageable given time and system limitations for this project. The PITCHf/x data does not include height, so additional data was downloaded from Baseball Prospectus to supplement the dataset with the pitcher's height for each record. A subset of the data was taken to eliminate irrelevant variables, including the 11 variables shown in Table 1 below. 13,095 incomplete records were removed, leaving 335,633 records for **total pitches** thrown. These are likely due to the occasional malfunction by the PITCHf/x technology, but the deleted records represented only a small portion of the dataset and their removal was not expected to have a significant effect on the results. The "total pitches" dataset was used to examine perceived velocity and downward plane. A second dataset was also created with the same observations, but including only **batted balls**. The "batted balls" dataset contained 60,515 observations and was used to analyze exit velocity and launch angle. An exploratory data analysis and correlational analysis were performed on the data.

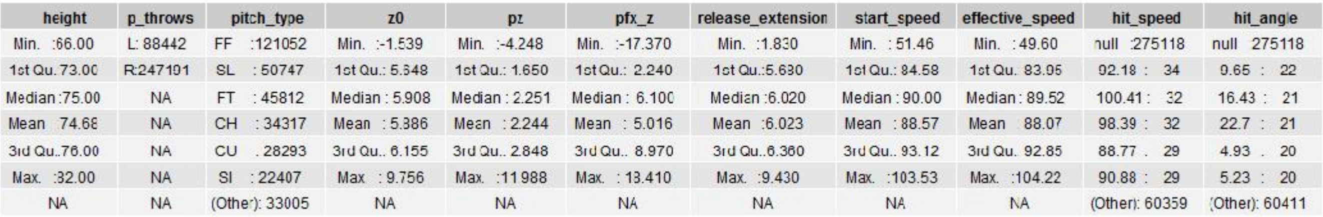
Table 1: Variable Definitions



Variable	Description	Type	Example
height	Height of the pitcher (inches)	Ratio	75
p_throws	Hand with which the pitcher throws	Nominal	R
pitch_type	Type of pitch thrown	Nominal	FF
z0	Height of the pitch measured at the initial point (feet)	Ratio	6.371
pz	Height of the pitch as it crossed the front of home plate (feet)	Ratio	4.578
pfx_z	Spin-induced vertical movement of the pitch between release point and home plate (inches)	Ratio	9.89
release_extension	Distance between the pitching rubber and the pitcher's release point (feet)	Ratio	6.02
start_speed	Velocity of the pitch measured at the initial point (MPH)	Ratio	96.10
effective_speed	"Perceived velocity" or how fast the pitch appears to a hitter (MPH)	Ratio	94.74
hit_speed	Speed of a baseball after it is hit by a batter (MPH)	Ratio	67.17
hit_angle	Vertical angle at which the ball leaves the bat after being struck	Ratio	45.87

Descriptive statistics were calculated as the initial step in the exploratory analysis. The mean height in the total pitches dataset was 74.68 inches, with a minimum of 66 inches and a maximum of 82. The mean is used later in the analysis to classify pitchers as tall or short. The min and max values do not appear initially to be extreme enough to be considered outliers. The remaining values in Table 2 illustrate the average measurements for release height, height at home plate, spin-induced vertical movement, extension, velocity, and perceived velocity during the 2016 season. The minimum start_speed of 51.46 MPH is notable, but certainly possible, and did not appear to have a drastic effect on the mean so no further action was taken.

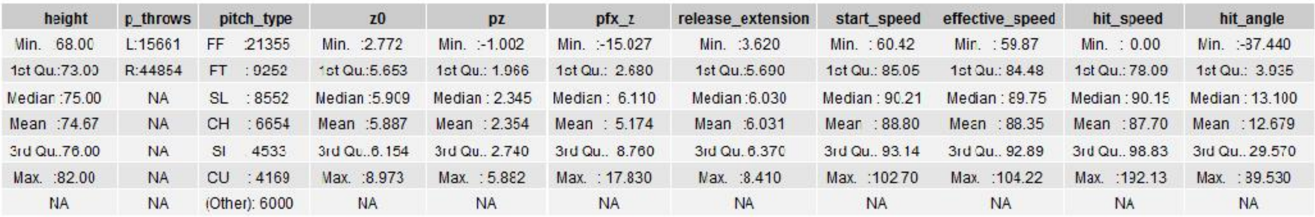
Table 2: Descriptive Statistics on all pitches thrown Apr-Jun 2016



height	p_throws	pitch_type	z0	pz	pfx_z	release_extension	start_speed	effective_speed	hit_speed	hit_angle
Min. :66.00	L: 88442	FF :121052	Min. :-1.539	Min. :-4.248	Min. :-17.370	Min. :1.830	Min. : 51.46	Min. : 49.60	null :275118	null 275118
1st Qu. 73.00	R:247191	SL : 50747	1st Qu.: 5.348	1st Qu.: 1.650	1st Qu.: 2.240	1st Qu.:5.630	1st Qu.: 84.58	1st Qu. 83.95	92.18 : 34	9.65 : 22
Median :75.00	NA	FT : 45812	Median : 5.908	Median : 2.251	Median : 6.100	Median :6.020	Median : 90.00	Median : 89.52	100.41 : 32	16.43 : 21
Mean 74.68	NA	CH : 34317	Mean : 5.386	Mean : 2.244	Mean : 5.016	Mean :6.023	Mean : 88.57	Mean 88.07	98.39 : 32	22.7 : 21
3rd Qu..76.00	NA	CU . 28293	3rd Qu.. 6.155	3rd Qu.. 2.848	3rd Qu.. 8.970	3rd Qu..6.360	3rd Qu.. 93.12	3rd Qu. 92.85	88.77 . 29	4.93 . 20
Max. :32.00	NA	SI : 22407	Max : 9.756	Max. :11.988	Max. : 13.410	Max. :9.430	Max. :103.53	Max. :104.22	90.88 : 29	5.23 : 20
NA	NA	(Other): 33005	NA	NA	NA	NA	NA	NA	(Other): 60359	(Other): 60411

The summary statistics for batted balls are shown below in Table 3. The average batted ball in 2016 travels at 87.70 MPH at an angle of 12.679 degrees. According to the Statcast Glossary, that is a line drive (Glossary). 10 degrees or less equates to a ground ball, between 10 and 25 degrees is line drive, between 25 and 50 degrees is a fly ball, and above 50 is a pop-up. This method was used later in the analysis to classify batted balls by result.

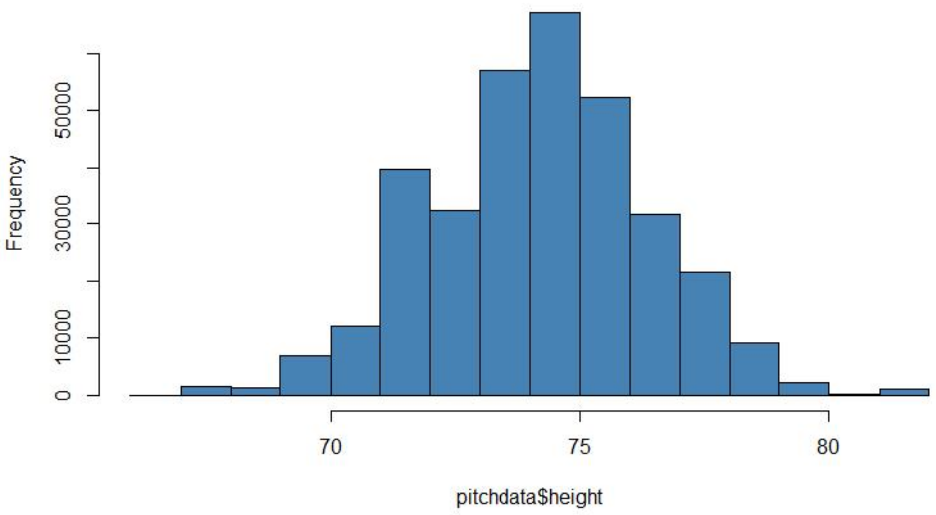
Table 3: Descriptive Statistics on batted balls Apr-Jun 2016



height	p_throws	pitch_type	z0	pz	pfx_z	release_extension	start_speed	effective_speed	hit_speed	hit_angle
Min. 68.00	L:15661	FF 21355	Min. :2.772	Min. :-1.002	Min. :-15.027	Min. :3.620	Min. : 60.42	Min. : 59.87	Min. : 0.00	Min. :-37.440
1st Qu.:73.00	R:44854	FT : 9252	1st Qu.:5.653	1st Qu.: 1.966	1st Qu.: 2.680	1st Qu.:5.690	1st Qu.: 85.05	1st Qu.: 84.48	1st Qu.: 78.09	1st Qu.: 3.935
Median :75.00	NA	SL : 8552	Median :5.909	Median : 2.345	Median : 6.110	Median :6.030	Median : 90.21	Median : 89.75	Median : 90.15	Median : 13.100
Mean :74.67	NA	CH : 6654	Mean :5.887	Mean : 2.354	Mean : 5.174	Mean :6.031	Mean : 88.80	Mean : 88.35	Mean : 87.70	Mean : 12.679
3rd Qu..76.00	NA	SI 4533	3rd Qu..6.154	3rd Qu.. 2.740	3rd Qu.. 8.760	3rd Qu. 6.370	3rd Qu.. 93.14	3rd Qu.. 92.89	3rd Qu.. 98.83	3rd Qu.. 29.570
Max. :82.00	NA	CU : 4169	Max. :8.973	Max. : 5.882	Max. : 17.830	Max. :8.410	Max. :102.70	Max. :104.22	Max. :192.13	Max. : 39.530
NA	NA	(Other): 6000	NA	NA	NA	NA	NA	NA	NA	NA

Histograms were produced to assess normality. Figure 1 below presents the histogram for height, featuring a reasonably normal distribution.

Figure 1: Height, total pitches thrown



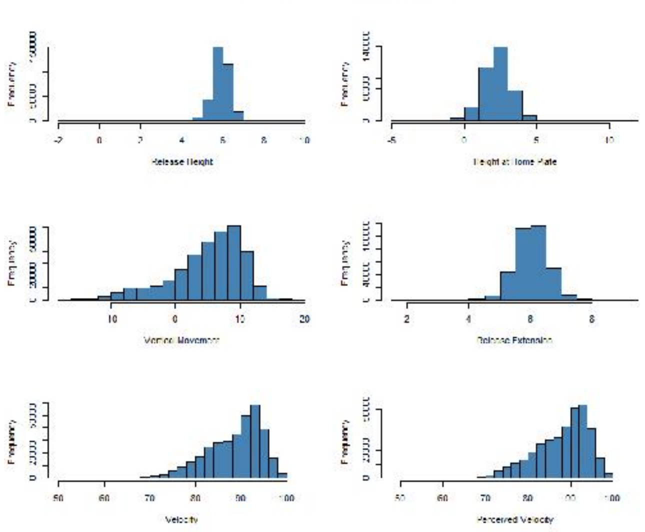


Figure 2: Histograms, all pitches thrown

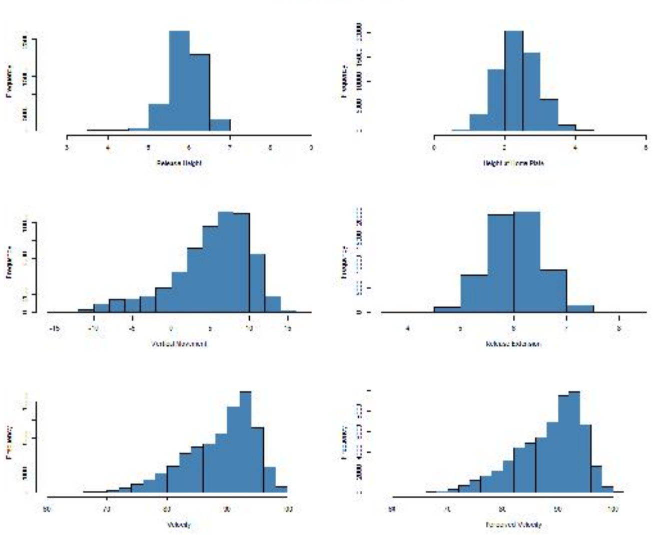


Figure 3: Histograms, batted balls only

Figures 2 and 3 above display histograms for the remaining variables in the datasets. For the most part, they appear to be normally distributed. There is some negative skewness, particularly in the velocity related fields, which is not surprising for the velocity related fields because most pitchers today throw higher than 90 MPH. In addition, the skewness and kurtosis values were calculated and are included below in Table 4. The values are a reasonable approximation to the values of 0 for skewness and 3 for kurtosis that represent a normal distribution, with the exception of the high kurtosis for release height. This indicates that the distribution for release height has heavy tails and outliers may be present.

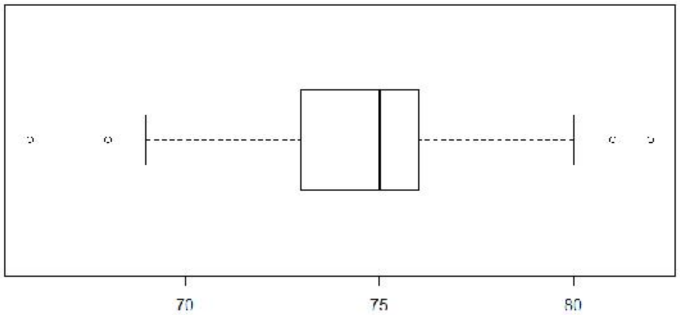
Table 4: Normality Assessment

Skewness of 0 and Kurtosis of 3 = Normal Distribution

	All pitches thrown		**Batted balls only**	
Variable	**Skewness**	**Kurtosis**	**Skewness2**	**Kurtosis2**
height	-0.0433042	3.060171	-0.0527742	3.079439
z0 (release height)	-0.9767748	7.752579	-0.9797308	7.327957
pz (height at home plate)	-0.0513740	3.533902	-0.0036745	3.325341
pfx_z (vertical movement)	-0.9069047	3.400290	-0.9636664	3.742167
release_extension	-0.0037659	3.088692	0.0249473	3.030414
start_speed (velocity)	-0.7613934	3.138539	-0.8109702	3.305817
effective_speed (perceived velocity)	-0.7561361	3.111096	-0.8020936	3.280604

Boxplots were produced to further assess the possibility of outliers in the data. Figure 4 below indicates 4 potential outliers based on height.

Figure 4: Height, all pitches thrown



As suspected earlier, the boxplots show extreme outliers in release height. A few records appear to have values near or below zero. These may be technical anomalies, but they may also be valid observations from a "submarine" style pitcher who releases the ball near the ground. Overall, the dataset is so large that the small number of potential outliers does not have a significant impact.

Figure 5: Boxplots, all pitches thrown

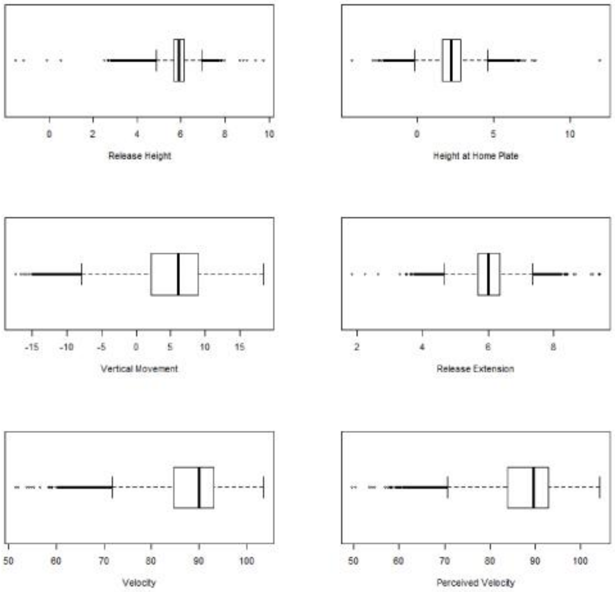
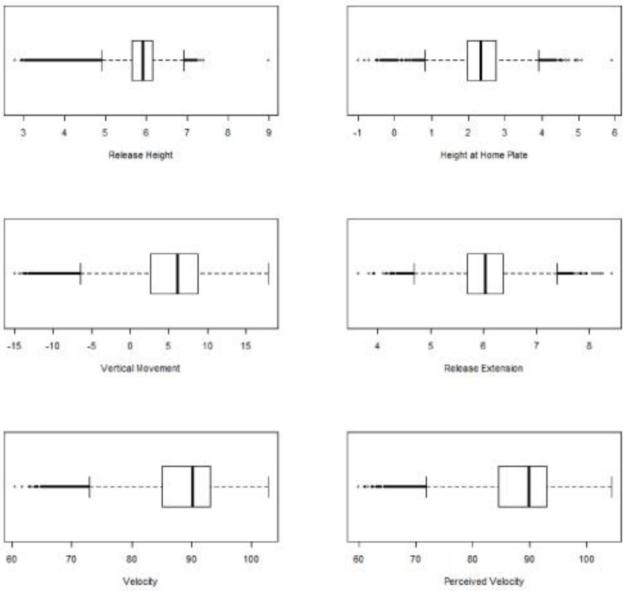


Figure 6: Boxplots, batted balls only

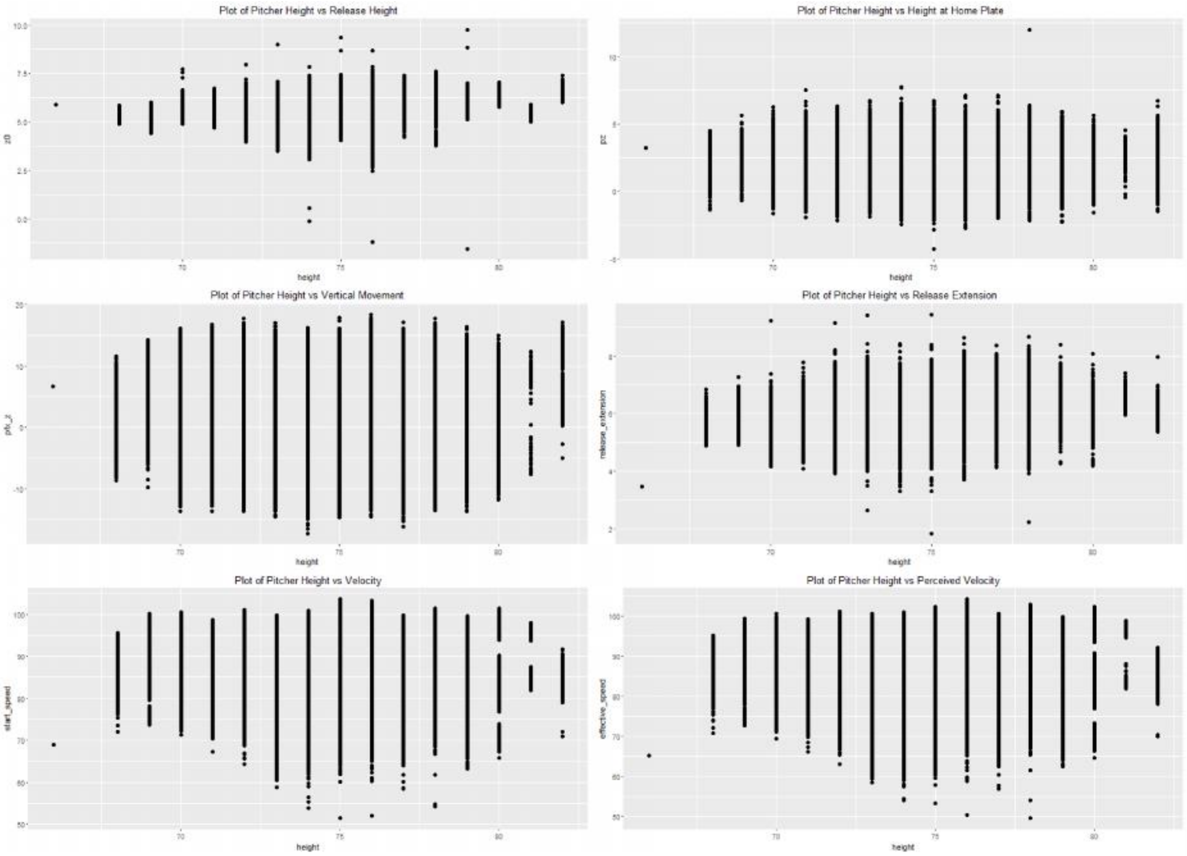


After assessing normality and outliers, a correlational analysis was performed on the data using both Pearson's product-moment correlation and Spearman's rho, along with scatterplots for visualization. Pearson correlation is commonly used to quantify linear relationships, while Spearman's rho depicts monotonic relationships. The data did not present clean linear relationships, so using both methods provides a more thorough analysis.

Results

Figure 7 illustrates the relationships between height and release height, height at home plate, spin-induced vertical movement, release extension, velocity, and perceived velocity. It is difficult to see any correlation amongst these variables. There is a barely distinguishable positive trend in Pitcher Height vs Release Height, driving the unsurprising inference that taller pitchers release the ball from higher point. The remaining plots each present an almost random distribution with no discernable pattern.

Figure 7: Scatterplots, all pitches thrown



Each of the above relationships was submitted through a correlational analysis, producing the values presented below in Table 5.

Table 5: Correlation with Height, all pitches thrown		
Variable	**Correlation (Pearson)**	**Correlation (Spearman)**
z0 (release height)	0.2962628 p-value < 2.2e-16	0.3182587 p-value < 2.2e-16
pz (height at home plate)	0.01734639 p-value < 2.2e-16	0.01564595 p-value < 2.2e-16
pfx_z (vertical movement)	-0.01673337 p-value < 2.2e-16	-0.009192637 p-value = 1.005e-07
release_extension	0.2581017 p-value < 2.2e-16	0.2676167 p-value < 2.2e-16
start_speed (velocity)	-0.003469711 p-value = 0.04442	0.03445507 p-value < 2.2e-16
effective_speed (perceived velocity)	0.02865105 p-value < 2.2e-16	0.06985312 p-value < 2.2e-16

As expected, release height exhibited a relatively strong correlation with height (r = 0.2963, rho = 0.3183). Release extension also produced a moderately strong correlation with height (r = 0.2581, rho = 0.2676). In both cases, Spearman’s rho is greater than Pearson’s r, indicating that the relationship is more monotonic than linear. While the correlation does not appear to be very strong, height is more positively correlated with perceived velocity (r = 0.0287, rho = 0.0699) than velocity (r = -0.0035, rho = 0.0345).

Height at home plate and vertical movement did not correlate strongly with height. Height at home plate is more dependent on the pitcher’s control and pitching style than height, while vertical movement from PITCHf/x is more dependent on the type of pitch thrown and shows a negative correlation to height. The attribute is defined as the vertical movement between release point and home plate, “as compared to a theoretical pitch thrown at the same speed with no spin-induced movement” (Fast). Essentially, it is the vertical movement driven by the pitch type, excluding the effects of gravity. This does not represent the “downward plane” discussed in the introduction. To create a simple approximation of the downward plane, a new variable was created as the difference between release height and height at home plate. The new variable, downward plane, yielded a much stronger correlation to height (r = 0.1165, rho = 0.1100).

Next, the relationship between height and pitching effectiveness, in terms of exit velocity, and launch angle, was examined. Correlations between exit velocity and height, release height, downward plane, release extension, and perceived velocity are presented in Table 6.

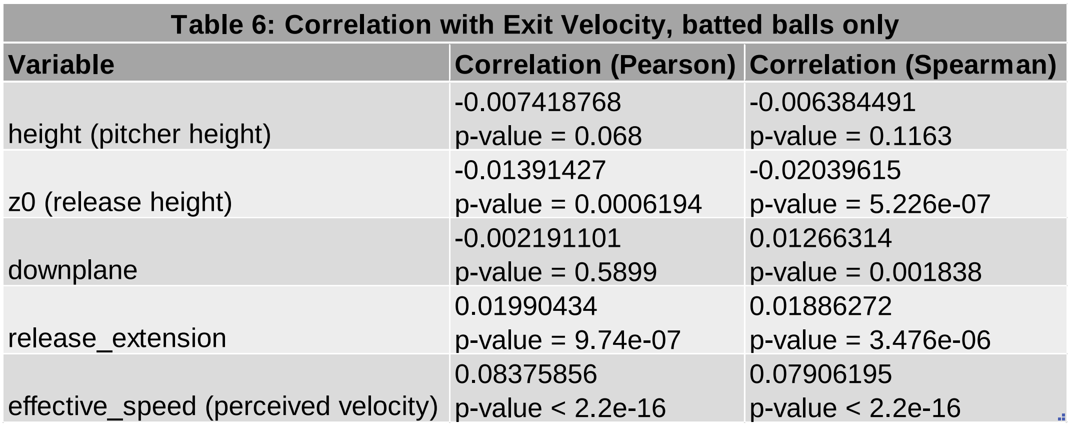


Table 6: Correlation with Exit Velocity, batted balls only

Variable	Correlation (Pearson)	Correlation (Spearman)
height (pitcher height)	-0.007418768 p-value = 0.068	-0.006384491 p-value = 0.1163
z0 (release height)	-0.01391427 p-value = 0.0006194	-0.02039615 p-value = 5.226e-07
downplane	-0.002191101 p-value = 0.5899	0.01266314 p-value = 0.001838
release_extension	0.01990434 p-value = 9.74e-07	0.01886272 p-value = 3.476e-06
effective_speed (perceived velocity)	0.08375856 p-value < 2.2e-16	0.07906195 p-value < 2.2e-16

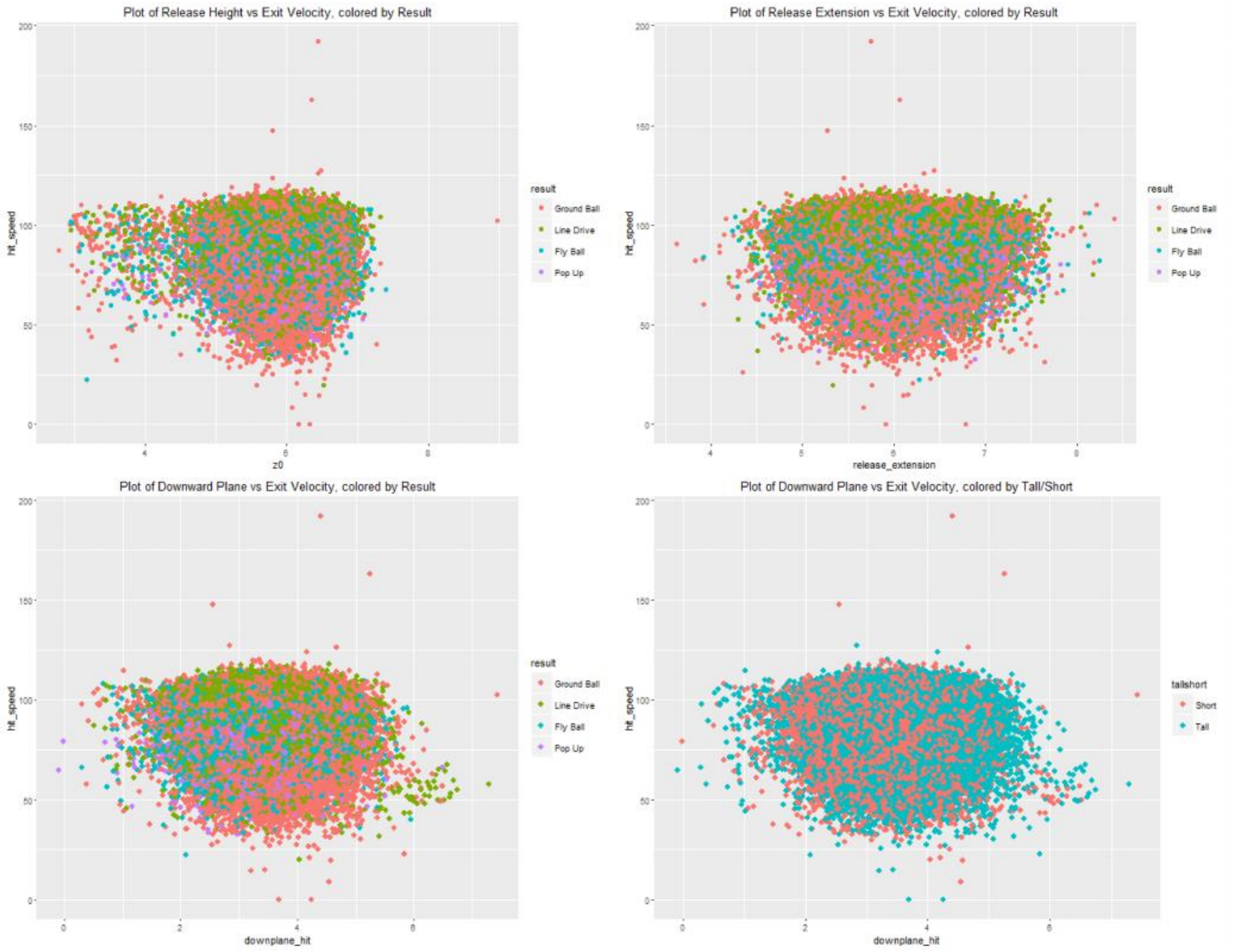
None of the variables display a strong correlation to exit velocity. There is no significant evidence of a negative correlation between downward plane and exit velocity (r = -0.0022, rho = 0.0013). There is actually a positive correlation between perceived velocity and exit velocity (r = 0.0838, rho = 0.0791).

The same variables were compared to launch angle, and the correlations are presented below in Table 7. In this case, downward plane is negatively correlated with launch angle (r = -0.1483, rho = -0.1574). On the other hand, perceived velocity produced a fairly weak positive correlation to launch angle (r = 0.0367, rho = 0.0386).

Table 7: Correlation with Launch Angle, batted balls only		
Variable	**Correlation (Pearson)**	**Correlation (Spearman)**
height (pitcher height)	0.01466573 p-value = 0.0003087	0.01261429 p-value = 0.001915
z0 (release height)	0.0191838 p-value = 2.364e-06	0.02026509 p-value = 6.178e-07
downplane	-0.1482963 p-value < 2.2e-16	-0.1574195 p-value < 2.2e-16
release_extension	0.01353231 p-value = 0.0008716	0.01091115 p-value = 0.007272
effective_speed (perceived velocity)	0.03666825 p-value = 0.04442	0.03859196 p-value < 2.2e-16

To reveal additional insight, observations were classified as tall or short based on whether the pitcher's height was above or below the mean height of 74.68. They were also classified by their result based on the launch angle specifications mentioned in the Methods section. The scatterplots in Figure 8 use these classifications to display relationships between exit velocities and launch angle.

Figure 8: Exit Velocity Scatterplots, batted balls only



The hypotheses that taller pitchers induce lower exit velocities and launch angles via a sharper downward plane stipulates that a weakly hit ground ball is the desired result. In the plot of release height vs exit velocity, this would manifest itself in higher number of ground balls in the lower right corner of the plot. However, the desired pattern is not present in Figure 8. There does appear to be a larger number of ground balls near a release height of 6 feet, but that value is hardly above the mean release height of 5.887 feet. Likewise, the plot of release extension vs exit velocity offers no evidence that taller pitchers who achieve greater extension see any benefit in effectiveness. The plot of downward plane vs exit velocity says little about exit velocity but shows that a higher downward plane leads to more ground balls than fly balls. The final plot displays the same data classified by tall/short rather than result and reveals that taller pitches dominate the right half of the plot.

Figure 9: Launch Scatterplots, batted balls only

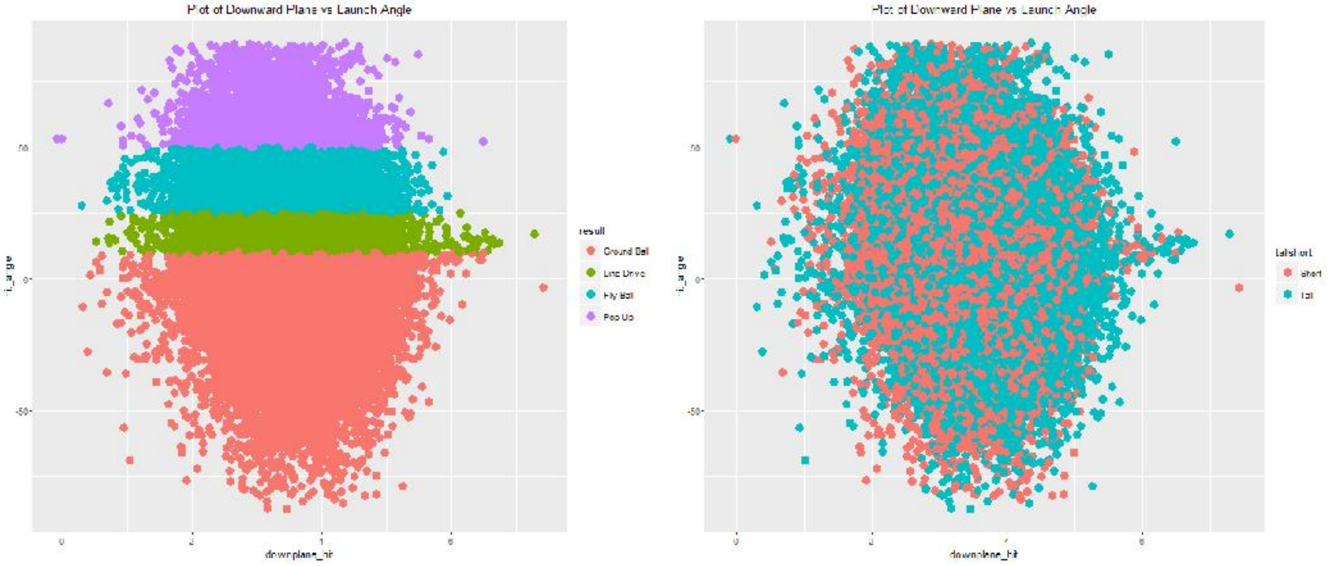


Figure 8 displays the relationship between downward plane and launch angle, classified by both result and tall/short. Once more, it is clear that taller pitchers are likely to produce a greater downward plane. It also appears that batted balls are more likely to be ground balls as downward plane increases.

Implications

The hypothesis that there would be a positive correlation between height and perceived velocity was true, though not to the degree that was expected. The correlation was rather weak, indicating that taller pitchers do not gain as much as common belief suggests in perceived velocity. The expectation that there would be a positive correlation between height and downward plane also proved to be true. Taller throwers can leverage their length to generate a more difficult angle for the hitter. In addition, the notion that there would be negative correlations between perceived velocity and both exit velocity and launch angle was demonstrably false. Both correlations were positive, offering no support of the theory that tall pitchers are able to induce weaker and less harmful contact by the batter. Finally, the hypothesis that there would be negative correlations between downward plane and both exit velocity and launch angle was partly true. While there was no support showing that a greater downward plane led to a lower exit velocity, there was a negative correlation between downward plane and launch angle. In other words, pitchers are more likely to produce a lower launch angle as downward plane increases.

The findings presented in this report can help reshape how Major League Baseball teams scout and search for pitchers. The analysis showed little correlation between height and perceived velocity, so teams seeking hard-throwing pitchers should not search on the basis of height. Taller pitchers may not be more likely throw the ball any harder than shorter pitchers, but their ability to create a greater downward plane from the release point to home plate may allow them to generate ground balls more frequently. Consequently, teams looking for an effective ground-ball-inducing pitcher may benefit from investigating taller pitchers.

This study did not distinguish between starting pitchers and relief pitchers. Often, these two types of pitchers have very different styles and goals and it may be wise to examine these relationships in that context. Furthermore, the data did not include any past seasons. Expanding the dataset would allow a more accurate examination of trends over time. The PITCHf/x data used in the study is more concentrated on the trajectory and spin of the ball than simple measures like height. Therefore, there may be confounding factors with some of the variables used in the analysis. There are also other fields that may prove enlightening on this topic, such as the swing and miss rate.

The correlational analysis was limited by its simplicity. A more comprehensive review or removal of outliers may have improved the model, though it is not likely with a dataset this large. A more complex analysis may look to build a predictive model for exit velocity and launch angle based on height and other related attributes, or compare tall and short pitchers to determine if there is a significant difference in certain areas.

Appendix A: References

Arthur, R. (2016, April 13). The New Science Of Hitting. Retrieved July 10, 2016, from http://fivethirtyeight.com/features/the-new-science-of-hitting/

Baseball Prospectus | Active Players by Year. (2016). Retrieved July 10, 2016, from http://www.baseballprospectus.com/sortable/extras/active_players.php?this_year=2016

Baseball Prospectus | Glossary. (2016). Retrieved July 10, 2016, from http://www.baseballprospectus.com/glossary/

Bryant, H. (2015, April 27). As athletes get bigger, they look less and less like us. Retrieved July 10, 2016, from http://espn.go.com/mlb/story/_/id/12751620/for-major-league-baseball-pitchers-bigger-better

Chicago Cubs 2016 Draft Results. (2016). Retrieved July 10, 2016, from http://chicago.cubs.mlb.com/team/draft.jsp?c_id=chc

Fast, M. (2007, August 02). Glossary of the Gameday pitch fields. Retrieved July 10, 2016, from https://fastballs.wordpress.com/2007/08/02/glossary-of-the-gameday-pitch-fields/

Glossary. (2016). Retrieved July 10, 2016, from http://m.mlb.com/glossary/statcast/

Greenberg, G. P. (2010). Does a Pitcher's Height Matter? Retrieved July 10, 2016, from http://sabr.org/research/does-pitcher-s-height-matter

Marchi, M., & Albert, J. (2013). Analyzing baseball data with R. Boca Raton, FL: CRC Press.

Rymer, Z. D. (2013, May 13). Do Taller Pitchers Throw Harder Than Average? Retrieved July 10, 2016, from http://bleacherreport.com/articles/1645950-do-taller-pitchers-throw-harder-than-average

Statcast Search. (2016). Retrieved July 10, 2016, from https://baseballsavant.mlb.com/statcast_search

Appendix B: R Code

```
#PREDICT 456 Sports Performance Analysis Section 55 Summer 2016
#Christopher Anderson
#Assignment #1

library(moments)
library(ggplot2)
library(gridExtra)
library(pitchRx)
library(Hmisc)

pitchdata <- read.csv("savant_data.csv", header = T, sep = ",")
hitdata <- read.csv("savant_data2.csv", header = T, sep = ",")

# Subset the data to include only relevant columns (348728 obs. of 11 variables)
pitchdata <- subset(pitchdata,
select=c("height","p_throws","pitch_type","z0","pz","pfx_z","release_extension",

"start_speed","effective_speed","hit_speed","hit_angle"))
hitdata <- subset(hitdata,
select=c("height","p_throws","pitch_type","z0","pz","pfx_z","release_extension",

"start_speed","effective_speed","hit_speed","hit_angle"))

# Remove records with missing data
pitchdata <- na.omit(pitchdata) # 335633 obs. of  11 variables (all pitches
thrown); 13095 removed
hitdata <- na.omit(hitdata) # 60515 obs. of  11 variables (batted balls); 288213
removed

# Examine structure of data and summary statistics
str(pitchdata)
head(pitchdata)
tail(pitchdata)
grid.table(summary(pitchdata))
dev.off()
str(hitdata)
head(hitdata)
tail(hitdata)
grid.table(summary(hitdata))
dev.off()

# Exploratory histograms and normality assessment for all pitches dataset
hist(pitchdata$height, main = "Figure 1: Height, total pitches thrown", col =
"steelblue")
skewness(pitchdata$height)
kurtosis(pitchdata$height)
par(mfrow = c(3,2), oma=c(0,0,2,0))
```

```
hist(pitchdata$z0, xlab = "Release Height", main = "", col = "steelblue")
hist(pitchdata$pz, xlab = "Height at Home Plate", main = "", col = "steelblue")
hist(pitchdata$pfx_z, xlab = "Vertical Movement", main = "", col = "steelblue")
hist(pitchdata$release_extension, xlab = "Release Extension", main = "", col =
"steelblue")
hist(pitchdata$start_speed, xlab = "Velocity", main = "", col = "steelblue")
hist(pitchdata$effective_speed, xlab = "Perceived Velocity", main = "", col =
"steelblue")
title("Figure 2: Histograms, all pitches thrown", outer=TRUE)
par(mfrow = c(1,1))
skewness(pitchdata$z0)
kurtosis(pitchdata$z0)
skewness(pitchdata$pz)
kurtosis(pitchdata$pz)
skewness(pitchdata$pfx_z)
kurtosis(pitchdata$pfx_z)
skewness(pitchdata$release_extension)
kurtosis(pitchdata$release_extension)
skewness(pitchdata$start_speed)
kurtosis(pitchdata$start_speed)
skewness(pitchdata$effective_speed)
kurtosis(pitchdata$effective_speed)

# Exploratory histograms and normality assessment for batted ball data
hist(hitdata$height, main = "Figure 1: Height, batted balls only", col =
"steelblue")
skewness(hitdata$height)
kurtosis(hitdata$height)
par(mfrow = c(3,2), oma=c(0,0,2,0))
hist(hitdata$z0, xlab = "Release Height", main = "", col = "steelblue")
hist(hitdata$pz, xlab = "Height at Home Plate", main = "", col = "steelblue")
hist(hitdata$pfx_z, xlab = "Vertical Movement", main = "", col = "steelblue")
hist(hitdata$release_extension, xlab = "Release Extension", main = "", col =
"steelblue")
hist(hitdata$start_speed, xlab = "Velocity", main = "", col = "steelblue")
hist(hitdata$effective_speed, xlab = "Perceived Velocity", main = "", col =
"steelblue")
title("Figure 3: Histograms, batted balls only", outer=TRUE)
par(mfrow = c(1,1))
skewness(hitdata$z0)
kurtosis(hitdata$z0)
skewness(hitdata$pz)
kurtosis(hitdata$pz)
skewness(hitdata$pfx_z)
kurtosis(hitdata$pfx_z)
skewness(hitdata$release_extension)
kurtosis(hitdata$release_extension)
skewness(hitdata$start_speed)
kurtosis(hitdata$start_speed)
```

```
skewness(hitdata$effective_speed)
kurtosis(hitdata$effective_speed)

# Check outliers using boxplots
boxplot(pitchdata$height, main = "Figure 4: Height, all pitches thrown", horizontal
= TRUE)
par(mfrow = c(3,2), oma=c(0,0,2,0))
boxplot(pitchdata$z0, xlab = "Release Height", main = "", horizontal = TRUE)
boxplot(pitchdata$pz, xlab = "Height at Home Plate", main = "", horizontal = TRUE)
boxplot(pitchdata$pfx_z, xlab = "Vertical Movement", main = "", horizontal = TRUE)
boxplot(pitchdata$release_extension, xlab = "Release Extension", main = "",
horizontal = TRUE)
boxplot(pitchdata$start_speed, xlab = "Velocity", main = "", horizontal = TRUE)
boxplot(pitchdata$effective_speed, xlab = "Perceived Velocity", main = "",
horizontal = TRUE)
title("Figure 5: Boxplots, all pitches thrown", outer=TRUE)
par(mfrow = c(1,1))

par(mfrow = c(3,2), oma=c(0,0,2,0))
boxplot(hitdata$z0, xlab = "Release Height", main = "", horizontal = TRUE)
boxplot(hitdata$pz, xlab = "Height at Home Plate", main = "", horizontal = TRUE)
boxplot(hitdata$pfx_z, xlab = "Vertical Movement", main = "", horizontal = TRUE)
boxplot(hitdata$release_extension, xlab = "Release Extension", main = "",
horizontal = TRUE)
boxplot(hitdata$start_speed, xlab = "Velocity", main = "", horizontal = TRUE)
boxplot(hitdata$effective_speed, xlab = "Perceived Velocity", main = "", horizontal
= TRUE)
title("Figure 6: Boxplots, batted balls only", outer=TRUE)
par(mfrow = c(1,1))

# Scatterplots to examine relationships between variables for all pitches thrown
data
plot1 <- ggplot(data=pitchdata, aes(x=height, y=z0)) + geom_point(size=2) +
                  ggtitle("Plot of Pitcher Height vs Release Height")
plot2 <- ggplot(data=pitchdata, aes(x=height, y=pz)) + geom_point(size=2) +
                  ggtitle("Plot of Pitcher Height vs Height at Home Plate")
plot3 <- ggplot(data=pitchdata, aes(x=height, y=pfx_z)) + geom_point(size=2) +
                  ggtitle("Plot of Pitcher Height vs Vertical Movement")
plot4 <- ggplot(data=pitchdata, aes(x=height, y=release_extension)) +
geom_point(size=2) +
                  ggtitle("Plot of Pitcher Height vs Release Extension")
plot5 <- ggplot(data=pitchdata, aes(x=height, y=start_speed)) + geom_point(size=2)
+
                  ggtitle("Plot of Pitcher Height vs Velocity")
plot6 <- ggplot(data=pitchdata, aes(x=height, y=effective_speed)) +
geom_point(size=2) +
                  ggtitle("Plot of Pitcher Height vs Perceived Velocity")
grid.arrange(plot1, plot2, plot3, plot4, plot5, plot6, ncol=2)
```

```
# Correlations with all pitches thrown data
cor.test(pitchdata$height, pitchdata$z0)
cor.test(pitchdata$height, pitchdata$z0, method="spearman")
cor.test(pitchdata$height, pitchdata$pz)
cor.test(pitchdata$height, pitchdata$pz, method="spearman")
cor.test(pitchdata$height, pitchdata$pfx_z)
cor.test(pitchdata$height, pitchdata$pfx_z, method="spearman")
cor.test(pitchdata$height, pitchdata$release_extension)
cor.test(pitchdata$height, pitchdata$release_extension, method="spearman")
cor.test(pitchdata$height, pitchdata$start_speed)
cor.test(pitchdata$height, pitchdata$start_speed, method="spearman")
cor.test(pitchdata$height, pitchdata$effective_speed)
cor.test(pitchdata$height, pitchdata$effective_speed, method="spearman")

# Create new variable for "downward plane"
# Height at release minus height as the ball crosses home plate
downplane <- pitchdata$z0-pitchdata$pz
pitchdata <- data.frame(pitchdata[1:6], downplane, pitchdata[7:11])
downplane_hit <- hitdata$z0-hitdata$pz
hitdata <- data.frame(hitdata[1:6], downplane_hit, hitdata[7:11])

cor.test(pitchdata$height, pitchdata$downplane)
cor.test(pitchdata$height, pitchdata$downplane, method="spearman")

# Correlations with exit velocity
# Batted balls only data
cor.test(y = hitdata$hit_speed, x = hitdata$height)
cor.test(y = hitdata$hit_speed, x = hitdata$height, method="spearman")
cor.test(y = hitdata$hit_speed, x = hitdata$z0)
cor.test(y = hitdata$hit_speed, x = hitdata$z0, method="spearman")
cor.test(y = hitdata$hit_speed, x = hitdata$downplane_hit)
cor.test(y = hitdata$hit_speed, x = hitdata$downplane_hit, method="spearman")
cor.test(y = hitdata$hit_speed, x = hitdata$release_extension)
cor.test(y = hitdata$hit_speed, x = hitdata$release_extension, method="spearman")
cor.test(y = hitdata$hit_speed, x = hitdata$effective_speed)
cor.test(y = hitdata$hit_speed, x = hitdata$effective_speed, method="spearman")

# Correlations with launch angle
# Batted balls only data
cor.test(y = hitdata$hit_angle, x = hitdata$height)
cor.test(y = hitdata$hit_angle, x = hitdata$height, method="spearman")
cor.test(y = hitdata$hit_angle, x = hitdata$z0)
cor.test(y = hitdata$hit_angle, x = hitdata$z0, method="spearman")
cor.test(y = hitdata$hit_angle, x = hitdata$downplane_hit)
cor.test(y = hitdata$hit_angle, x = hitdata$downplane_hit, method="spearman")
cor.test(y = hitdata$hit_angle, x = hitdata$release_extension)
cor.test(y = hitdata$hit_angle, x = hitdata$release_extension, method="spearman")
cor.test(y = hitdata$hit_angle, x = hitdata$effective_speed)
cor.test(y = hitdata$hit_angle, x = hitdata$effective_speed, method="spearman")

# Create new variable for "result"
# Classification of result as Ground ball, line drive, fly ball, or pop up
result <- cut(hitdata$hit_angle, c(-180,10,25,50,180), c("Ground Ball","Line
Drive","Fly Ball","Pop Up"))
hitdata <- data.frame(hitdata, result)
# Create new variable "tallshort"
# Classification of pitcher as tall or short based on mean height
tallshort <- cut(hitdata$height, c(60,74.67,90), c("Short","Tall"))
hitdata <- data.frame(hitdata[1], tallshort, hitdata[2:13])

# Scatterplots of interesting relationships between variables for all batted balls
data
plot1 <- ggplot(data=hitdata, aes(x=z0, y=hit_speed, colour=result)) +
geom_point(size=2) +
  ggtitle("Plot of Release Height vs Exit Velocity, colored by Result")
plot2 <- ggplot(data=hitdata, aes(x=release_extension, y=hit_speed, colour=result))
+ geom_point(size=2) +
  ggtitle("Plot of Release Extension vs Exit Velocity, colored by Result")
plot3 <- ggplot(data=hitdata, aes(x=downplane_hit, y=hit_speed, colour=result)) +
geom_point(size=2.5) +
  ggtitle("Plot of Downward Plane vs Exit Velocity, colored by Result")
plot4 <- ggplot(data=hitdata, aes(x=downplane_hit, y=hit_speed, colour=tallshort))
+ geom_point(size=2.5) +
  ggtitle("Plot of Downward Plane vs Exit Velocity, colored by Tall/Short")
grid.arrange(plot1, plot2, plot3, plot4, ncol=2)

plot1 <- ggplot(data=hitdata, aes(x=downplane_hit, y=hit_angle, colour=result)) +
geom_point(size=4) +
  ggtitle("Plot of Downward Plane vs Launch Angle")
plot2 <- ggplot(data=hitdata, aes(x=downplane_hit, y=hit_angle, colour=tallshort))
+ geom_point(size=4) +
  ggtitle("Plot of Downward Plane vs Launch Angle")
grid.arrange(plot1, plot2, ncol=2)

# Save datasets for future use
write.csv(pitchdata, file = "pitchdata.csv")
write.csv(hitdata, file = "hitdata.csv")
```